STA 602 Lab 4

Yutong Shao

Jan.31, 2023

```
library(ggplot2)
```

Review

```
# generate k samples from normal gamma dist.
# normal gamma is the joint prior distribution
rNG=function(k, m0, n0, a, b){
  lambda=rgamma(k, a, b)
  mu=rnorm(k,m0,sqrt(1/(n0*lambda)))
  return(rbind(mu, lambda))
}
rNG(3, 0, 1, 1, 1)
```

```
##             [,1]      [,2]       [,3]
## mu     0.5635126 1.6787928 -0.1084703
## lambda 1.3044965 0.8079633  3.7497950
```

```
# three samples
```

Here how to update the parameters

```
# posterior distribution
NG.update=function(m0,n0,a,b, X){
  n=length(X)
  mX=sum(X)/n
  ssX=sum((X-mX)^2)
  m0.post=(n*mX+n0*m0)/(n+n0)
  n0.post=n+n0
  a.post=a+n/2
  b.post=b+(ssX+(n*n0)/(n+n0)*(mX-m0)^2)/2
  return(list(m0.post=m0.post,n0.post=n0.post,a.post=a.post,b.post=b.post))
}
```

If now the true mean is 1 and the true precision is 3, and our prior guess are 0 for the mean and 1 for the precision with psudo samples sizes equal to 1, let's see if we can recover the truth with different sample sizes

```r
N=1000

mu.true=1
lambda.true=3
m0=0
n0=1
lambda1=1
n1=1
a=n1/2
b=n1/(2*lambda1)

X = rnorm(N, mu.true, sd=sqrt(1/lambda.true))
update=NG.update(m0, n0, a, b, X)

mu.post=update$m0.post
lambda.post=update$a.post/update$b.post

c(mu.true,mu.post)
```

```
## [1] 1.000000 1.031934
```

```r
c(lambda.true,lambda.post)
```

```
## [1] 3.000000 2.832067
```

Now a function to sample from the posterior

```r
rNG.post=function(k,m0,n0,a,b,X){
  update=NG.update(m0, n0, a, b, X)
  m0.post=update$m0.post
  n0.post=update$n0.post
  a.post=update$a.post
  b.post=update$b.post
  return(rNG(k,m0.post,n0.post,a.post,b.post))
}

rNG.post(10,m0, n0, a, b, X)
```

```
##            [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]     [,8]
## mu     1.039682 1.022959 1.019253 1.021596 1.024975 1.027027 1.012051 1.049423
## lambda 2.872952 2.681676 2.621392 2.563211 2.918141 2.651104 2.795900 2.662733
##            [,9]    [,10]
## mu     1.025709 1.024691
## lambda 3.001575 2.895565
```

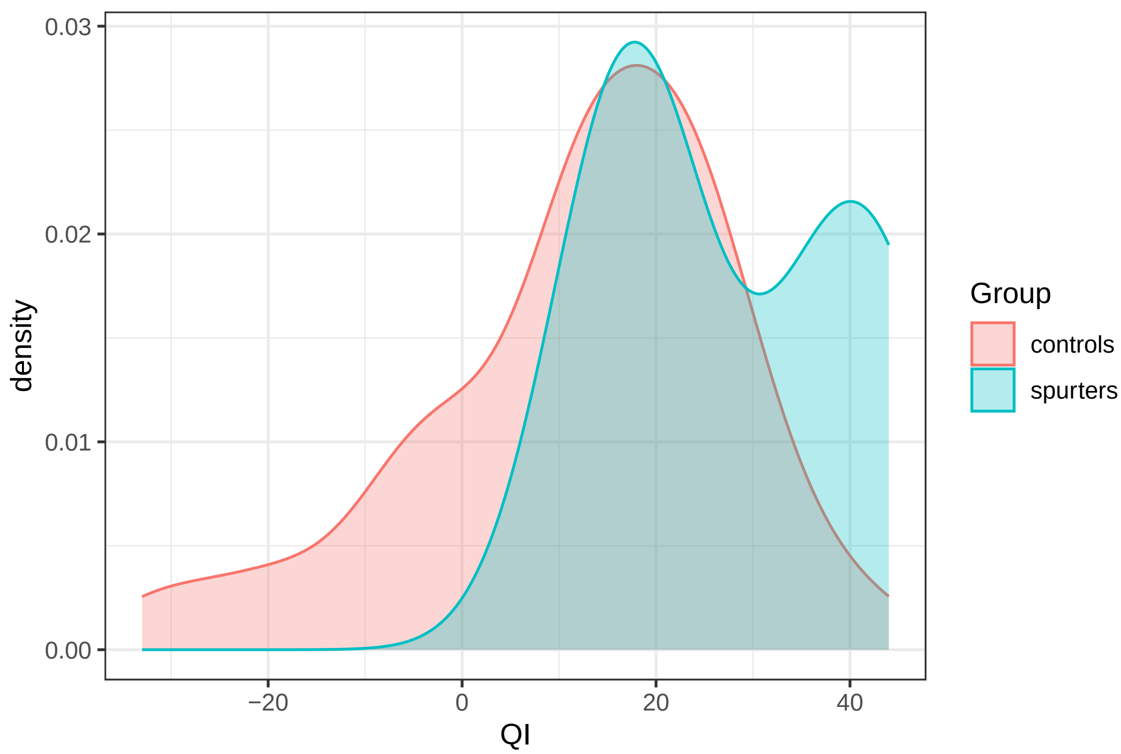
Do teacher's expectations influence student achievement?

Students had an IQ test at the begining and end of a year; the data is the difference in IQ score. **20% of the students** were randomly chosen; their teacher was told they were "spurters" (high performers)

```r
spurters = c(18, 40, 15, 17, 20, 44, 38)
controls = c(-4, 0, -19, 24, 19, 10, 5, 10,
             29, 13, -9, -8, 20, -1, 12, 21,
             -7, 14, 13, 20, 11, 16, 15, 27,
             23, 36, -33, 34, 13, 11, -19, 21,
             6, 25, 30,22, -28, 15, 26, -1, -2,
             43, 23, 22, 25, 16, 10, 29)
```

Task 1: Plot histograms for the change in IQ score for the two groups. Report your findings.

```r
# just combine the above two data frames and assign their labels, whether
# 'spurter' or 'controls'
data=data.frame(QI=c(spurters,controls),Group=as.factor(c(rep("spurters",7),rep("controls",48))))

ggplot(data,aes(x=QI,col=Group,fill=Group))+
  geom_density(alpha=0.3)+
  theme_bw()
```



Task 2: How strongly does this data support the hypothesis that the teachers expectations caused the spurters to perform better than their classmates?

Let's use a normal model:

$$X_1, \dots, X_{n_s} \mid \mu_s, \lambda_s^{-1} \overset{iid}{\sim} \text{Normal}(\mu_S, \lambda_S^{-1})$$

$$Y_1, \dots, Y_{n_C} \mid \mu_c, \lambda_c^{-1} \overset{iid}{\sim} \text{Normal}(\mu_C, \lambda_C^{-1}).$$

We are interested in the difference between the means—in particular, is $\mu_S > \mu_C$?

We can answer this by computing the posterior probability that $\mu_S > \mu_C$:

$$\mathbb{P}[\mu_S > \mu_C | x_{1:n_S}, y_{1:n_C}] = \mathbb{E}[\mathbf{1}_{\mu_S > \mu_C} | x_{1:n_S}, y_{1:n_C}].$$

Let's assume independent Normal-Gamma priors:

$$\text{spurters: } (\mu_S, \lambda_S) \sim \text{NormalGamma}(m, c, a, b)$$
$$\text{controls: } (\mu_C, \lambda_C) \sim \text{NormalGamma}(m, c, a, b)$$

Subjective choice:

Remark: **One could choose these parameters based on data. But they should NOT entirely depend on data, because we don't know the data before setting prior distribution.**

- $\mu_0 = 0$ Don't know whether students will improve or not, on average
- $n_0 = 1$ Weakly informative prior; pseudo sample size equal to 1/10
- $n_1 = 1$ Weakly informative prior; pseudo sample size equal to 1/10
- $\lambda_1 = 1/10^2$ We expect the standard deviation to be around 10.
- Thus, $a = 1/2$, $b = 50$.

```
sd(controls)
```

```
## [1] 16.27288
```

```
m = 0
c = 1
a = 1/2
b = 50
```

Now let's sample from the posterior distributions.

```
k = 10000
spurters.sampled =
  rNG.post(k, m, c, a, b, spurters)
controls.sampled =
  rNG.post(k, m, c, a, b, controls)
```

Using the Monte-Carlo approximation

$$\mathbb{P}(\mu_S > \mu_C \mid x_{1:n_S}, y_{1:n_C}) = \mathbb{E}[\mathbf{1}_{\mu_S > \mu_C} | x_{1:n_S}, y_{1:n_C}] \approx \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\mu_S^{(i)} > \mu_C^{(i)}},$$

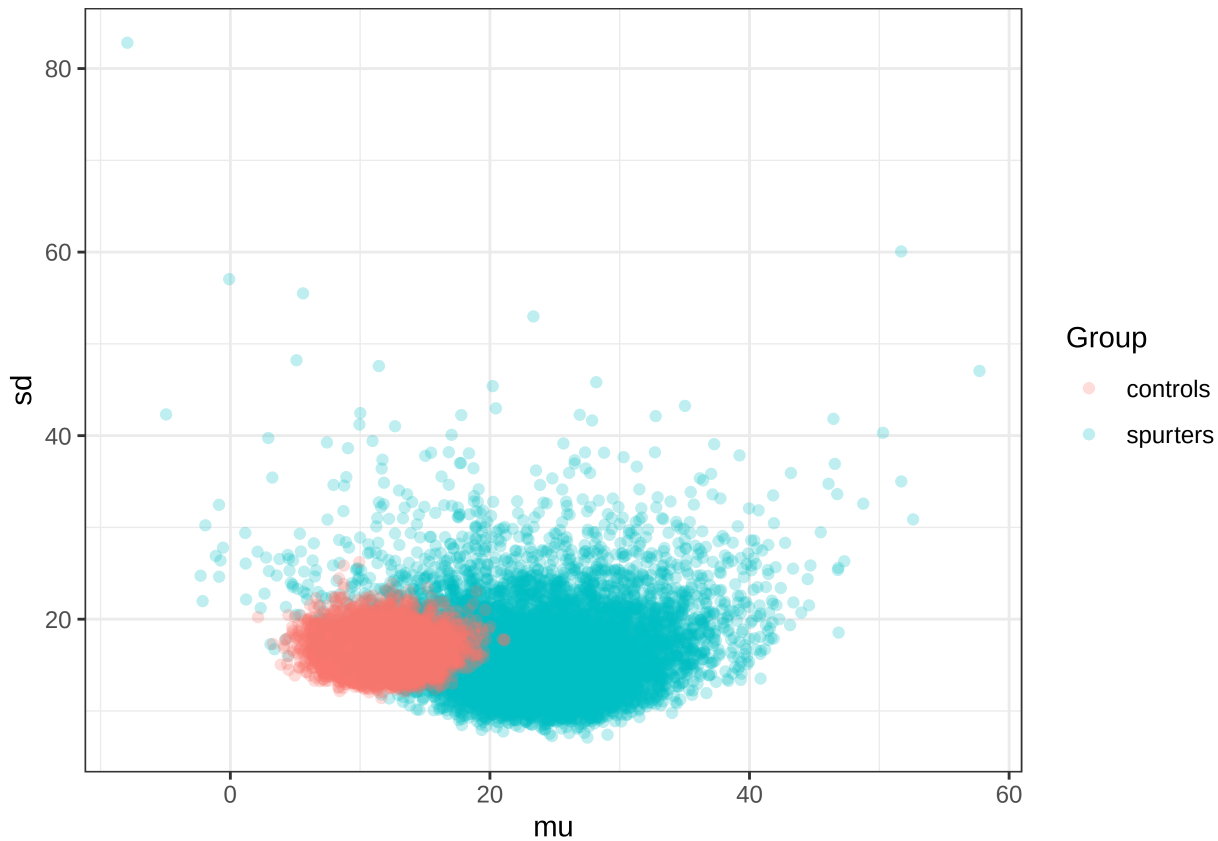
we find

```
mean(spurters.sampled["mu",]>controls.sampled["mu",])
```

```
## [1] 0.9712
```

Task 3: Provide a scatterplot of samples from the posterior distributions for the two groups. What are your conclusions?

```r
dd=data.frame(mu=c(spurters.sampled[1,],controls.sampled[1,]),
              sd=c(spurters.sampled[2,]^(-1/2),controls.sampled[2,]^(-1/2)),
              Group=as.factor(c(rep("spurters",k),rep("controls",k))))

ggplot(data=dd,aes(x=mu,y=sd,col=Group))+
  geom_point(alpha=0.25)+
  # xlim(0,50)+ylim(0,50)+
  theme_bw()
```




```r
sd(spurters.sampled[2,]^(-1/2))
```

```
## [1] 4.777612
```

```r
sd(controls.sampled[2,]^(-1/2))
```

```
## [1] 1.701347
```

Task 4: Compute the probability that

$$\mathbb{P}(\mu_S > \mu_C \mid x_{1:n_S}, y_{1:n_C}) = \mathbb{E}[\mathbf{1}_{\mu_S>\mu_C}|x_{1:n_S}, y_{1:n_C}] \approx \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\mu_S^{(i)}>\mu_C^{(i)}},$$

and interpret the posterior probability.

```
k = 10000
spurters.sampled =
  rNG.post(k, m, c, a, b, spurters)
controls.sampled =
  rNG.post(k, m, c, a, b, controls)

mean(spurters.sampled["mu",]>controls.sampled["mu",])
```

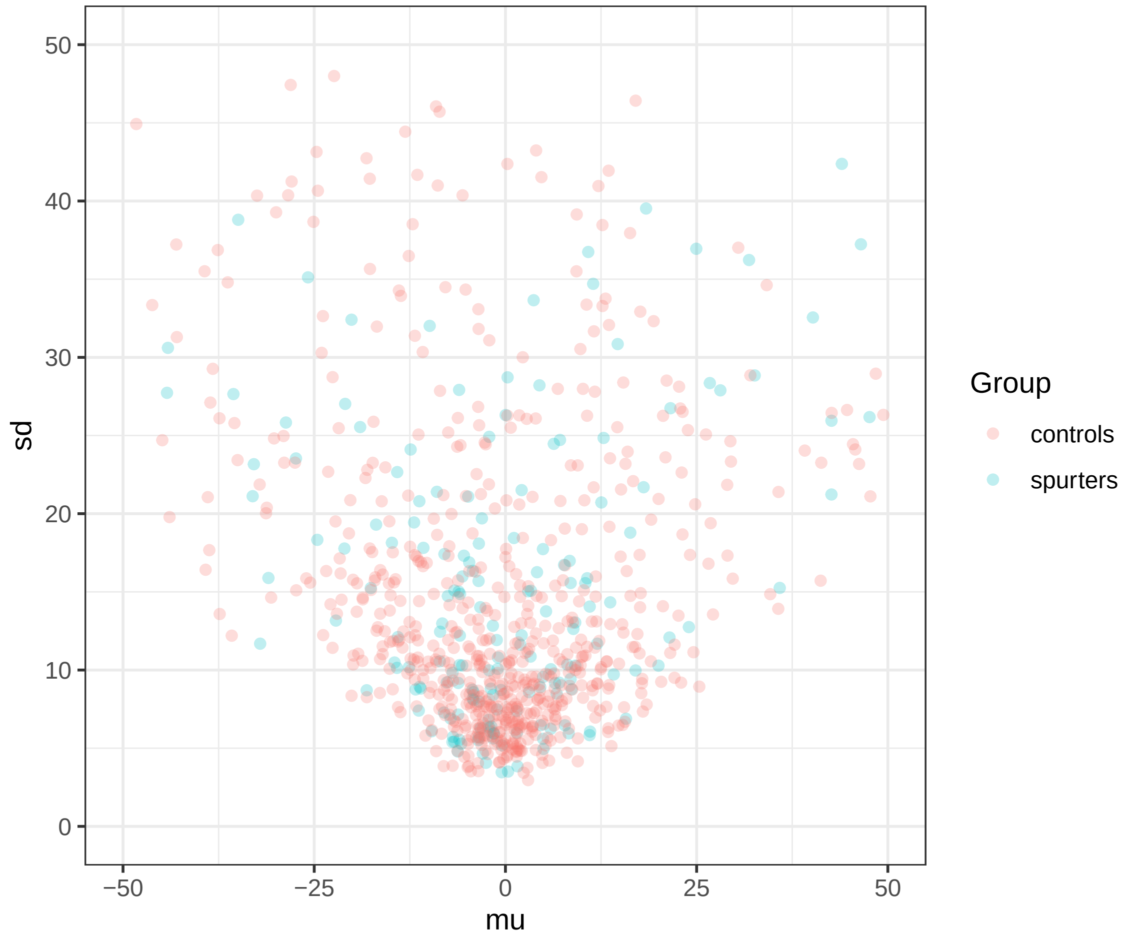
```
## [1] 0.9706
```

The posterior probability I computed above means on average, around 97% of the spurters' change in IQ are higher than that of the controls, which further indicates teacher's expectation do has influences on students' IQ.

Task 5: Replicate Figure 3

```
sim <- 1000
spurters_sim <- sim * 0.2
ctrl_sim <- sim * 0.8

# print(m)

spurters_prior =
  rNG(spurters_sim, m, c, a, b)

ctrl_prior =
  rNG(ctrl_sim, m, c, a, b)

df <-  data.frame(mu=c(spurters_prior[1,], ctrl_prior[1,]),
              sd=c(spurters_prior[2,]^(-1/2), ctrl_prior[2,]^(-1/2)),
              Group=as.factor(c(rep("spurters",spurters_sim),
                                rep("controls",ctrl_sim))))

ggplot(data=df,aes(x=mu,y=sd,col=Group))+
  geom_point(alpha=0.25)+
  xlim(-50,50)+ylim(0,50)+
  theme_bw()
```

```
## Warning: Removed 181 rows containing missing values (`geom_point()`).
```



From the graph we can conclude that mean and variance of both controls and spurters follows Normal-Gamma distribution. The majority two groups are concentrated near mean=0, which is conform to our assumptions that students in both groups has no significant changes before the treatment. So it is a reasonable assumption.